The International Journal of Human Resource Management,
Vol. 21, No. 15, December 2010, 2813–2830

Routledge
Taylor & Francis Group

360 degree feedback: how many raters are needed for reliable ratings on the capacity to develop competences, with personal qualities as developmental goals?

Rainer Hensel[a]*, Frans Meijers[a], Rien van der Leeden[b] and Joseph Kessels[c]

[a]*Research Group of Professional Development in Vocational and Organisational Learning, The Hague University of Professional Development, The Hague, The Netherlands;* [b]*Institute of Psychology, Methodology and Statistics Unit, Leiden University, Leiden, The Netherlands;* [c]*Department of Organisational Psychology and Human Resource Development, Twente University, Enschede, The Netherlands*

360 degree feedback is a widely used technique in the area of strategic human resource management (SHRM) and strategic human resource development (SHRD). The reliability of 360 degree feedback on the capacity to develop personal qualities has been investigated. This study shows to what extent the number of raters is related to an increasing reliability and an enhancement of correlation between supervisor and peer ratings. Ten raters are needed to reach a satisfying reliability level of 0.7 for the rating of the capacity to develop personal qualities, while six raters are needed for a reliability level of 0.7 with regard to the rating of motivation to develop these qualities. The use of two or three peer raters, as is common in the daily HRM/HRD practice, results in low reliability levels and in low agreement between supervisor and peer ratings. These results imply that 360 degree feedback is more useful in a personal growth system than in an administrative system, where the outcomes of the feedback are considered to be objective representations of work behaviour. Further implications for the SHRM/SHRD practice, especially concerning the development of competences, with personal qualities as developmental goals, are discussed.

Keywords: competence development; personal qualities; reliability; rating; supervisor peer agreement; 360 degree feedback

Introduction

One of the core activities of strategic human resource development (SHRD) is to develop human qualities that are of critical importance for the successful realization of the organizational strategy. These human qualities, especially those related to personality traits measured by the Five Factor model, frequently form the essence of a competency profile (Furnham 2008, p. 318). Competence is defined as a cluster of related knowledge, skills, attitudes, values and personal qualities requiring a person to be successful performing a group of related tasks (Blanchard and Thacker 2007). Competences have unique characteristics or qualities that are difficult to copy. To make them applicable for SHRD purposes it should be possible to train and develop competences by SHRD programs (Hamel and Prahalad 1994) Competence development is considered to be a key element of any HRD policy (Walton 1999; Blanchard and Thacker 2007). There is an increasing tendency for SHRD professionals to focus on personal qualities (Blanchard and Thacker 2007, p. 20). This tendency is supported

*Corresponding author. Email: r.w.hensel@hhs.nl

ISSN 0958-5192 print/ISSN 1466-4399 online

DOI: 10.1080/09585192.2010.528664
http://www.informaworld.com

by studies showing a relation between personal qualities and a broad range of aspects of organizational effectiveness (Salgado 1997; Anderson and Viswesvaran 1998; Barrick, Mount and Judge 2001; Arthur, Bennet, Edens and Bell 2003).

In the daily practice of HRD, monitoring of progress and systematically providing feedback on the development of personal qualities gets very little attention (Bassi and Van Buren 1999; Walton 1999; Blanchard and Thacker 2007). Feedback on behavioural change, especially repeated feedback, has a significant positive effect on training effectiveness (Brett and Atwater 2001). Feedback in general is an important variable increasing the effectiveness of training and development (Houston 1990). Studies have shown that feedback not only enhances personal performance in many areas, but also increases the intrinsic motivation of employees (Hackman and Oldman 1975, 1980). Employees experienced feedback as a helpful HRD instrument for planning developmental goals and behavioural change (McCarthy and Garavan 1999). However, the meta study by Fried and Ferris (1987) has shown that the experienced meaningfulness of feedback was a mediating variable in the relationship between feedback and performance. Feedback should provide employees with meaningful information concerning their behavioural change, feedback can only be considered as meaningful when reliable. Even negative feedback enhances the effectiveness of personal change when it provides employees with information about discrepancies between a desired standard and their current state (Atwater, Roush and Fishthal 1995; Walker and Smither 1999). Moreover, feedback on strengths and weaknesses improves team performance (Lassiter 1996; Martineau 1998). But if the discrepancy between self-assessments and the received feedback is greater than expected, emotions, like anger and discouragement, can have a negative effect on behavioural change that is stronger when the feedback relates to personal qualities (Brett and Atwater 2001). This might be related to the fact that ratings of personal qualities are far more complex than ratings on instrumental skills or performance (Viswesvaran, Ones and Schmidt 1996; Arvey and Murphy 1998). Feedback is considered to be important for the enhancement of self knowledge, this could be due to the fact that self-ratings are problematic. Serious problems have been reported concerning the use of self-ratings: leniency, unreliability, bias and affects by numerous factors such as age, gender, personality, and self esteem (Hoffman, Nathan and Holden 1991; Yammarino and Atwater 1997; Beehr, Ivanitskaya, Hansen, Erofeev and Gudanowski 2001). It seems to be difficult to rate one's own abilities or effectiveness of work behaviour in a reliable and valid way. Therefore, 360 degree feedback is considered to be important for the enhancement of self-knowledge.

The use of unreliable ratings on performance, abilities and developmental capacities can cause serious problems. The malfunctioning of ratings assessing employee abilities caused major distrust and moral problems in organizations, leading to effectiveness problems (Fahr, Cannella and Bedeian 1991; Andrews 1997; Bettenhausen and Fedor 1997; Drenth 1998; Scholtes 1999; Gray 2002). According to Drenth (1998), a lack of statistic accuracy, objectivity, reliability and validity will damage the fairness perception of employees. Reliability and trustworthiness of feedback is especially important when personal qualities have to be developed, unreliable feedback on personal qualities provokes strong negative emotions (Brett and Atwater 2001). As mentioned above, personal qualities form an important part of the concept competency (Furnham 2008, p. 318). Although research has shown that ratings used for appraisal and assessment can be problematic, the application of single item ratings on complex work behaviour enjoys great popularity (Rasch 2004). Ratings made by supervisors/managers show considerable variation. The study by Gwynne (2002) has shown that variation was related to the work ethics of raters, rather than to variation of characteristics of the employees being judged. Variables concerning the rating system could be seen as the

major cause for error (Deming 1986). Additional problems with ratings by individuals are a *lack of accuracy* caused by latent motives or values, *inflated ratings*, *leniency*, *halo and horn effects* and *less variability* (Bernardin and Pence 1980; Ilgen and Feldman 1983; Landy and Farr 1983; Banks and Murphy 1985; Longenecker, Sims and Gioia 1987; Fahr, Cannella and Bedeian 1991; Murphy and Cleveland 1991; Bretz, Milkovich and Read 1992; Harris 1994; Murphy and Cleveland 1991; Drenth 1998). Leniency was related to accountability and defined as: 'the need to justify the rating to the employee being rated or to significant others in a face to face situation' (Roch and McNall 2007). Leniency can be caused by accountability because raters feel pressure to please the other, or to avoid mistakes that would cause embarrassment by the employee being rated. Advantages related to political power struggles in the social network of employees and managers seem to cause low agreement levels within organizations using ratings, especially when personal qualities have to be assessed (Kenny, Albright, Malloy and Kashy 1994). The results of a study by Kenny et al. (1994) seem to indicate that power struggles within the organization can damage the validity and reliability of the rating.

Problems concerning self-ratings and rating by one individual stimulated many organizations to use 360 degree feedback. The central assumption in using 360 degree feedback is that aggregated scores of several raters will result in a more accurate representation of the actual work behaviour (Robinson and Robinson 1989). 360 degree feedback is often called a multi-source feedback. It is a widely used technique to improve the reliability and validity of ratings of employee abilities or performances (London and Smither 1995; Church and Bracken 1997; Toegel and Conger 2003; Society for Human Resource Management & Personnel Decisions International 2000). However, the use of 360 degree feedback is frequently criticized. Metastudies showed that the inter-rater agreement in multi-source rating of all sort of performances is low (Conway and Huffcutt 1997). Correlations between the assessment of managers and subordinates were 0.14. Correlations between supervisors and peers were a bit higher at 0.34. Even measuring correlations within the same rating source did not result in higher correlations (Greguras and Robie 1998; Mount, Judge, Scullen, Sytsma and Hezlett 1998). Measuring methods contributed to a great extent to rating variance (Mount et al. 1998; Scullen, Mount and Judge 2003). When raters are instructed that their judgment is for developmental purposes, agreement levels for self-supervisor and self-peer rise, but stay relatively low (London and Beatty 1993; Waldman and Atwater 1998). Research on the validity of multi-source ratings by using externally validated criteria showed non-significant correlations (Van Hooft, van der Flier and Minne 2006).

Disappointing reliability and validity levels could be due to the fact that the average number of peer raters used for 360 degree feedback is too low (Van Hooft et al. 2006). Although many researchers state that a minimum of three to five peer raters should be used in combination with one supervisor rating for reliable 360 degree feedback (Bracken 1994; Antonioni 1996; Pollack and Pollack 1996; Lepsinger and Lucia 1997), it seems to be common practice to use the rating of one supervisor and only two or three peer raters when 360 degree feedback is applied (Rasch 2004; Van Hooft, van der Flier and Minne 2006). As far as we know, no studies are available that investigate to what extent reliability levels rise when raters are added to the rating system. Subsequently, no studies seem to be available that illustrate whether the use of the advised number of three to five raters will lead to a satisfactory reliability level. According to Nunnally (1987) a satisfactory level of reliability is 0.7.

In this paper we try to determine how many raters are needed for reliable 360 degree feedback to judge the capacity to develop personal qualities, closely related to personality traits measured by the Five Factor model of personality. Focus is laid on personal qualities because competency profiles are intuitive taxonomies in modern business language

representing the Five Factor personality factors lexically (Furnham 2008, p. 318). Therefore participants of this study were selected with learning goals that could be related to the Five Factor model of personality. Two additional reasons for focusing on personal qualities are: (1) ratings on personal qualities are qualified as more unreliable than ratings on instrumental behaviours (Viswesvaran 1996; Arvey and Murphy 1998); and (2) unreliable ratings on personal qualities seem to evoke strong negative emotions (Deming 1992; Brett and Atwater 2001), increasing the chance of the occurrence of moral and distrust problems in organizations (Fahr et al. 1991; Andrews 1997; Bettenhausen and Fedor 1997; Drenth 1998; Scholtes 1999; Gray 2002).

Validity and reliability problems seem to play an important role when studying problems concerning single and multi source ratings. In this study only the reliability of 360 degree feedback on the effectuation of the development of competences is investigated. This study focuses on single item multi source 360 degree feedback. The reason we concentrate on a single item is its widespread use, especially for administrative purposes (George 1994; Drenth 1998; Mani 2002; Rasch 2004). We would expect that the quality of measurement to improve by using multi-item scales. However, Jellema (2003), found disappointing reliability levels for a multi-item instrument for 360 degree feedback on HRD training effects. Adding items to scales is one possibility in enhancing reliability of feedback on training effects; another possibility is to add raters. This study focuses on the effect on the reliability of a single item measure, when individual raters are added to the rating system. Subsequently, we only focus on 'off the job' training activities because 65% of HRD activities take place off the job (Berghenegouwen and Mooijman 2010).

The following research questions are studied for the application of 360 degree feedback on the capacity to develop competences, with personal qualities as training goals: (1) What is the increase of reliability of a single item measure when peer raters are added in the process of rating?; (2) How many peer raters are required to reach a satisfactory reliability level of 0.7?; and (3) What is the effect of adding peer raters in the process of rating on the correlation between aggregated peer ratings and the supervisor rating?

Method

Participants and procedure

Data were obtained from a SHRD training program. Participants of this training were professionals and managers from a wide range of different organizations. The training was executed in small training groups with 12 participants. This setting was considered to be appropriate because the influence of political power struggles on the ratings was expected to be very low. The SHRD program is a competence-based communication. Central topics of the training are the development and training of multiple leadership styles using the contingency leadership concept; developing multiple regulative organizational roles for group decision-making and problem-solving; and training and development of social skills and conflict-solving skills. In general, the training has a strong focus on giving and receiving personal feedback and developing competences that are supportive to the variety of an organization's strategies.

The training procedure is standardized with a total duration of 12 days plus eight evening sessions, distributed over a period of six months in four periods of three successive days, plus two evenings. Towards the end of the training, particularly during the last six days and four evenings, responsibility for the learning process is delegated to the group members. Managers and professionals follow exactly the same program, but in separate groups.

At the end of the training sessions the training goals of all group members were collected, listed and distributed among the group members and the trainer. To be included in the study, participants were required to have training goals that could be linked to personality traits measured by the Five Factor model of personality. All of the participants could base their goal setting on a development assessment using the Dutch version of the Neo Pi-R test (McCrae and Costa 1989) measuring the Five Factor model of personality traits. Examples of training goals that can be linked to personality traits are: decreasing cognitive and affective rigidity during conflict; or developing a leadership style that is supportive for employees to develop talents. Examples of training goals that led to exclusion are; improving time management or the acquisition of instrumental skills to be used for appraisal reasons. Fifteen possible participants, who appeared to have training goals that could not be linked to the listed criteria, were excluded.

Participants were informed that the collected data would be used for research purposes. In addition they were asked to sign a letter of informed consent. Three possible candidates refused to sign and were excluded from this study. Data could be collected for 236 participants from 22 training groups. This resulted in 22 supervisor and 5192 peer ratings. Two people expressed the wish not to be rated. Thirty-four per cent of the sample is female, the average age is 38.6 (SD = 7.4). The data relating to 11 participants could not be collected because of absence on the day ratings were collected. All group members were trained in differentiating between observable behaviour and the interpretation of that behaviour, giving feedback on personal development based on observed behaviour and recognizing personal rating mistakes, like leniency, halo and horn effects. The use of dysfunctional political power influencing others was also discussed and group members were stimulated to express feedback whenever dysfunctional power use occurred. The reason for this is that training raters in the proper use of appraisal techniques has an important positive effect on the reliability of ratings (Woehr and Huffcut 1994; McEnery and Blanchard 1999; Bracken, Timmreck, Fleenor and Summers 2001).

Measures

For all participants ratings were collected on the capacity to effectuate the development of competences, as well as a rating of motivation to develop competences related to personal qualities. Single item multi-source measures were constructed for both variables with a maximum of eleven raters. Each group member, including the training supervisor, rated all other group members on both variables with a score ranging from one to 10. Training supervisors instructed the group members in the exact use of the scores and related criteria. For instance, concerning the achievement of personal training goals, a '1' indicated no lasting change at all; a '3' meant little change. A '5' meaning that a person has shown reasonable intention/effort to accomplish training goals, but could not accomplish an observable, lasting behavioural change. A '6' meaning that only limited lasting change has occurred, '7' indicating that a reasonable amount of change has occurred and a score of '10' was given if the change was very large in comparison to the observed-level at the beginning of the training program. Judgments were supposed to be based on concrete, well observable patterns of new behaviour, directly related to the underlying learning goals. During training, group members designed a special learning situation for each other where new behaviour patterns could be developed that were directly related to the underlying personal quality the trainee wished to see develop. Motivation was measured using the same method. A '1' meant that motivation was very low and absolutely insufficient for personal change, a '3' indicated a rating of low motivation. A '5' meant that motivation was reasonable but insufficient for personal change,

'6' was given if motivation was considered lower than average but just sufficient enough for personal change to occur. A '7' indicated higher motivation giving a good basis for personal change and a '10' meant very high motivation. All numbers below six would be given if the rater judged the other group member as externally motivated to participate in training. All numbers ranging from '1' to '10' could be used.

Statistical analyses

We focus upon the reliability of a single item multi-source rating of the capacity to develop competences and motivation. Multi-source, here, amounts to combining the ratings for a training group member, obtained from all of a number of other training group members on these two variables. Hence, an estimate of the reliability of the aggregated ratings for an individual was computed (sum-score or average). This reliability estimate was based on an approach from *generalizability* theory which comes close to computing the intra-class correlation in a three-level hierarchical data structure (cf. Hox 2002). In our case there are three different sources of variance: variability between ratings of the same individual; variability across the aggregated ratings of the individuals within each group; and variability between groups. Reliability estimates were obtained from the variance component estimates of a three-level model. Further details are described in the Appendix. The multilevel analysis program MLWiN version 2.02 (Rasbash, Steele, Browne and Prosser 2004) was used for model estimation. Reliability estimates were compared for the multi-source rating obtained from an increasing number of raters, ranging from two to 11. A reliability level of 0.7 or higher was considered to be a satisfying level of reliability (Nunnally 1978). To investigate the level of agreement between the supervisor of the training and the group members (further indicated as peers), correlations were computed between the aggregated ratings of peers and training supervisors. Ratings of training supervisors and aggregated ratings of peers were compared for an increasing number of peer raters, starting with two with a maximum of 11.

Results

Ratings of the capacity to develop competences

Using the maximum number of available raters resulted in a reliability level of 0.72. Ratings of motivation and the capacity to develop competences appeared to be highly correlated ($r = 0.61, p = 0.00$). For an increasing numbers of raters the obtained reliability estimate is presented in Table 1 below.

Table 1 clearly shows that the reliability level of ratings of the capacity to develop competences increases with an increasing number of raters. Using a small amount of raters, two or three, leads to insufficient reliability levels between 0.45 and 0.5. Using five raters results in a reliability level of 0.6. One needs 10 raters to reach the desired reliability level of 0.7. In Figure 1 these results are visualized.

Figure 1 illustrates the steady increase of the reliability when more than two raters are used. The strongest enhancement can be observed between two and six raters. When using more than six or seven raters the increase diminishes.

Using the maximum number of available raters (11) the correlation between supervisor ratings and the aggregated ratings on the capacity to develop competences is: 0.5

Table 1. Reliability estimates of the variable measuring the capacity to develop competences for increasing number of raters.

# Raters	2	3	4	5	6	7	8	9	10	11
Reliability	0.45	0.50	0.54	0.60	0.63	0.65	0.67	0.69	0.70	0.72

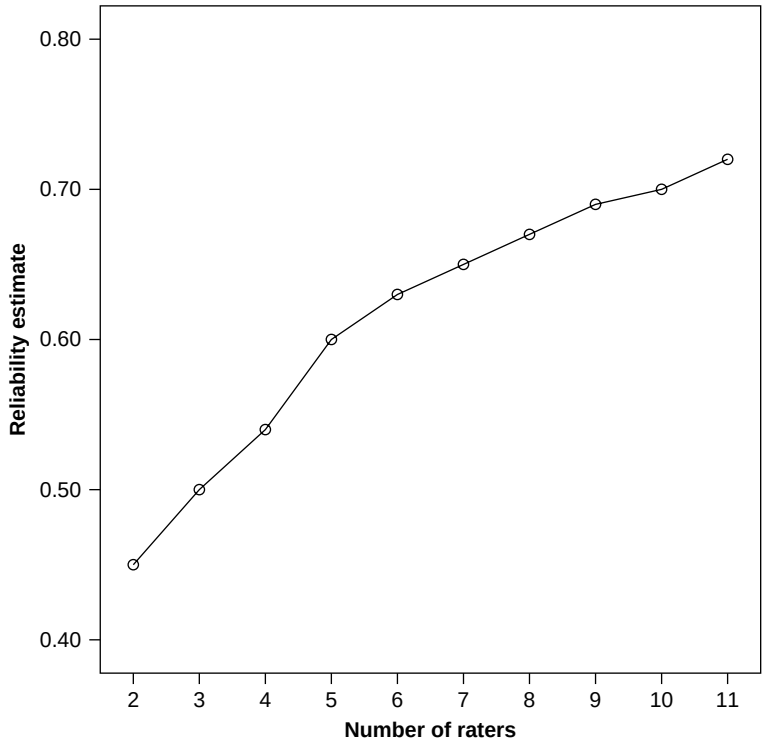

Figure 1. Reliability estimates of the variable measuring the capacity to develop competences for increasing number of raters.

($p = 0.00$). Correlations between supervisor and group member ('peers') ratings were calculated for an increasing number of peer raters, starting with two with a maximum of 11. Results are presented in Table 2 below.

Table 2 shows that the correlation increases with increasing number of raters. For two or three raters correlations are relatively low. A relatively high correlation (> 0.45) is obtained using six or more raters. To achieve a correlation of 0.5, 11 raters were necessary. In Figure 2 these results are visualized.

Figure 2 illustrates the increase of the correlation when more than two raters are used. The strongest increase can be observed between two and six raters. When using more than six raters the increase is less substantial.

Ratings of motivation to develop competences

Table 3 shows that the reliability level of the rating of motivation increases with an increasing number of raters. A small amount of raters, two or three, leads to a reliability

Table 2. Correlations between supervisor ratings and aggregated ratings of peers (training group members) for the variable measuring the capacity to develop competences for increasing number of raters.

# Raters	2	3	4	5	6	7	8	9	10	11
Reliability	0.31	0.35	0.40	0.43	.46	0.48	0.45	0.48	0.48	0.50
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

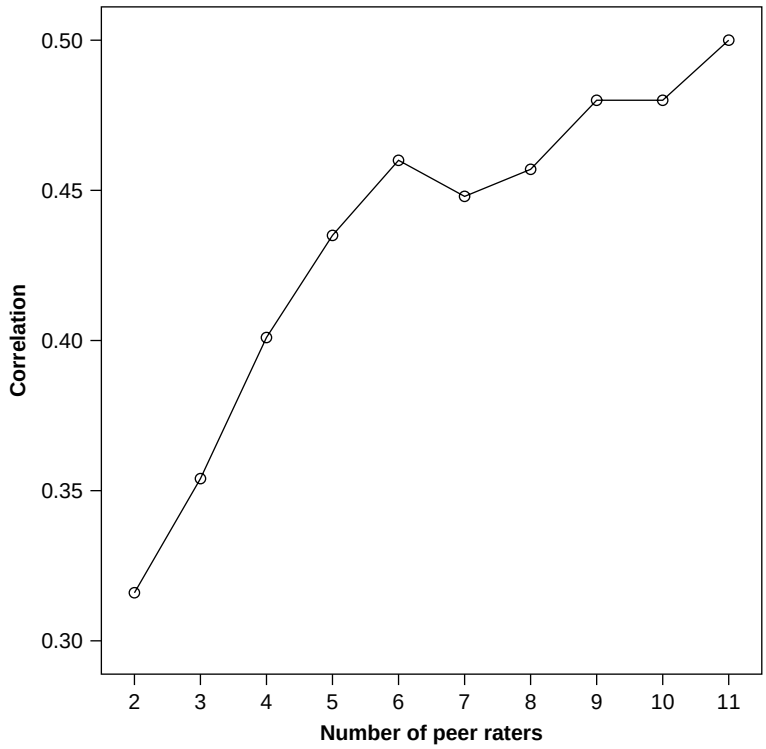

Figure 2. Correlations between supervisor ratings and aggregated ratings of peers (training group members) for the variable measuring the capacity to develop competences for increasing number of raters.

level of 0.47 or 0.49. One needs six raters to reach a satisfying reliability level of 0.7. In Figure 3 these results are visualized.

Figure 3 illustrates that the strongest enhancement can be observed between two and six raters. Using more than six or seven raters does not lead to a strong enhancement of reliability levels when motivation is rated.

Table 4 demonstrates that the correlation increases with an increasing number of raters when motivation is judged. A relatively high correlation (> 0.45) is obtained using six or more raters. After seven raters the reliability levels for motivation decrease.

Figure 4 is used to visualize the relationship between an increasing number of raters and the correlation between aggregated peer ratings and the supervisor's judgment on motivation. Although the line increases between two and seven raters, the line declines after seven raters making it difficult to observe a steady increasing or decreasing line.

Conclusion

This study has intentionally focused on feedback on the capacity to develop competences, with personal qualities as developmental goals. The development of competences should be

Table 3. Reliability estimates of the variable measuring motivation for increasing number of raters.

# Raters	2	3	4	5	6	7	8	9	10	11
Reliability	0.47	0.49	0.57	0.65	0.70	0.72	0.73	0.75	0.77	0.78

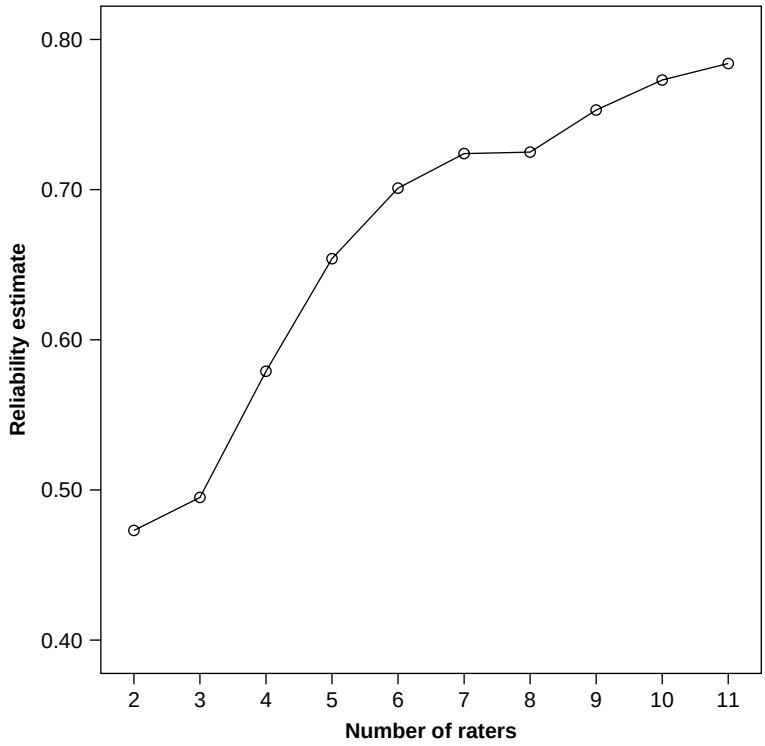

Figure 3. Reliability estimates of the variable measuring motivation for increasing number of raters.

an elementary part of a SHRD policy as it should deliver essential employee qualities to the organization to be effective. Personal qualities, especially those measured by the Five Factor model of personality, are related to a broad range of aspects of the organizational effectiveness (Salgado 1997; Anderson and Viswesvaran 1998; Barrick et al. 2001; Arthur et al. 2003). Results of this study show that an increasing number of raters leads to an enhancement of the reliability of peer ratings and the levels of agreement between supervisors and peers, when the effectiveness of competence development is rated. There is a clear relationship between an increasing number of peer ratings and rising reliability levels concerning the rating of the effectuation of learning goals by competence development. The same conclusions can be drawn concerning the relationship between the number of raters and the correlations between supervisor ratings and aggregated peer ratings. Higher correlations are achieved if the number of peer raters is increased. To reach a reliability level of 0.7 a large number of raters (at least 10) are needed. The use of two or three peer raters – which seems to be common practice in SHRD programs – results in a reliability level of 0.45 respectively 0.5, which cannot be considered as satisfying if one wants an accurate judgment. Unless one considers variation based on differences in perspective, as interesting and valuable for the learning process.

Table 4. Correlations between supervisor ratings and aggregated ratings of peers (training group members) for the variable measuring motivation for increasing number of raters.

# Raters	2	3	4	5	6	7	8	9	10	11
Reliability	0.37	0.42	0.43	0.42	.45	0.46	0.43	0.42	0.40	0.41
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

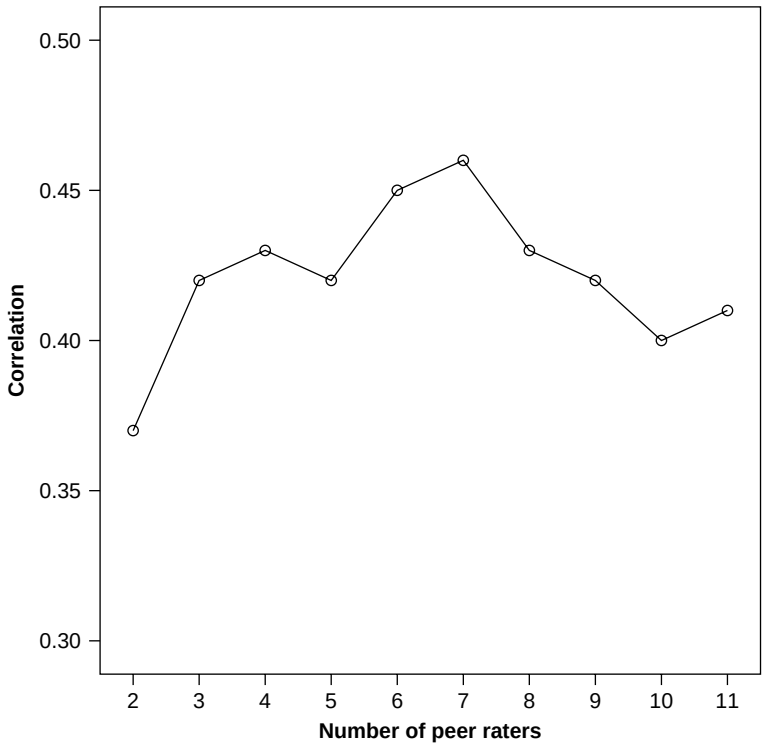

Figure 4. Correlations between supervisor ratings and aggregated ratings of peers (training group members) for the variable measuring motivation for increasing number of raters.

A stronger effect of increasing reliability levels can be observed concerning the reliability of the ratings of peers on *motivation* if compared to the peer ratings on the *effectiveness on competence development*. Here also, the reliability levels are quite unsatisfying: 0.47 (two raters), 0.49 (three raters), when two or three raters are used. Six raters are necessary for the desired reliability level of 0.7. The use of six raters does not seem to be realistic in the daily work flow.

Interpretation of the agreement of peer and supervisor ratings on *motivation* seems to be quite difficult. Although increasing correlations can be observed when the number of raters rises from two to seven, after this point (0.46) the correlations decrease, without reaching the low level computed for two raters (0.37). No clear conclusions can be drawn from these study results concerning the agreement between the ratings of peers and supervisors concerning motivation. A very plausible explanation could be derived from the fact that the supervisor of the training has given a lot of attention to the specification and commitment for achieving personal training goals. These efforts have not been done in the area of motivation. Differences in perspective on motivation between the supervisors and peers can easily be held responsible for these diffuse study results. The model of Latham and Locke (1991) has shown that goal specificity and goal commitment increases performance. Lack of consensus on perspectives to judge motivation can give a plausible explanation for the diffuse pattern concerning the rating agreement between supervisors and peers for motivation.

The results of this study support the underlying rationale of the application of 360 degree feedback. Adding raters leads to increasing reliability levels with each rater added to the rating system, when 360 degree feedback is used for competence development, with

personal qualities as developmental goals, but the number of raters must be high. As mentioned before, on average two or three raters are used when 360 degree is used for administrative reasons (Rasch 2004; Van Hooft et al. 2006).

Rising reliability levels highlight opportunities for monitoring training effects when competency, especially personal qualities, seems to be important. To create possibilities for reliable monitoring on competence development with personal qualities as developmental goals, organizations could strive for an organizational culture where honest and open feedback is available. Then a high number of raters is available and giving and receiving feedback in an open way seems to be part of the cultural values and norms within the organization. If this is the case, obtaining measures for feedback on HRD progress could be easily organized. When seven to eight raters are used, the reliability level exceeds 0.65, approaching the desired level of 0.7. Raters should be trained in the proper use of appraisal techniques as was the case with raters participating in this study. In general, training raters is a wise thing to do as it enhances the reliability levels of raters (Woehr and Huffcut 1994; McEnery and Blanchard 1999; Bracken et al. 2001). However, creating an open and honest culture is not an easy thing to accomplish. The appreciation of open and honest feedback is a personality trait (McCrae and Costa 1989; McCrae and John 1992). Personality traits are relatively stable, aggregated personality measures in a setting are significantly related to specific aspects of the organizational culture (Schneider 1987; Holland 1997). The work of Schneider (1987) in particular, has shown that people are attracted to, selected by and stay with organizations with internal cultures that fit their personality. When the shared value concerning the appreciation of honest and open feedback is low it will demand specific and complex leadership skills to shape the culture in the desired direction. This study has shown that designing a system with enough raters to ensure satisfying reliability and inter-rater levels is worthwhile. It leads to higher reliability levels than administrative procedures measuring training effectiveness by using one supervisor and two or three peer raters. This study has shown that an increasing number of raters increase the reliability levels of the rating on competence development. If one designs a learning culture, where judgments of a high number of raters are available, one creates the possibility for a broad multi-source rating that will lead to satisfying reliability levels. It should be accentuated that rating differences should still be accepted. Differences are based on a wide variety of points of view or perspectives and when properly used, differences enhance learning capacities by adding individual perspectives (Robinson and Robinson 1989). It could also lead to a culture where training effectiveness or behavioural change by personal growth is systematically monitored.

As feedback is important for the effectiveness of an SHRD policy the application of a systematic single item 360 degree feedback measurement could deliver valuable information enhancing feedback mechanisms. Although it seems to be time consuming, it should be highlighted that the lack of a systematic evaluation of training effect is considered to be a serious shortcoming of many HRD strategies (Gerber 1995; Bassi and Van Buren 1999; Walton 1999; Blanchard and Thacker 2007). As feedback is an important variable for competence development, the effectiveness of HRD policies should be monitored, but this aspect gets little attention in the daily practice of SHRD professionals (Bassi and Van Buren 1999; Walton 1999; Blanchard and Thacker 2007). The effectiveness of only 12% of the training programs, focusing on behaviour change, was evaluated (Bassi and Van Buren 1999). Reliable ratings on the effectuation of learning goals could be an interesting method to monitor HRD effectiveness and assess developmental capacities when personal qualities are important for the work.

If one seeks systematic evaluation the cost/benefit ratio seems to be reasonable as only one item is used, but more than eight raters (at least) are necessary. The use of a single item

measure, directly asking raters to judge the effectuation of competence development creates the possibility of evaluating training effects in a systematic way. It seems to be quite unsatisfying that only 12% of training activities concerning behavioural change are evaluated (Bassi and Van Buren 1999).

Implications for HRD systems for competence development

The results of this study may have a number of implications for the systems that are used for competence development. Two systems can be distinguished that are used for competence management: a supervisor directed/administrative system and a personal growth system (George 1994; Mani 2002). The major difference between the two systems is that differences in self-other and supervisor – peer ratings, caused by different perspectives, are accepted when the personal growth system is applied (Jellema 2000; Van Hooft et al. 2006). A planned dialogue to discuss and analyze differences is an essential part of the personal growth system. When the administrative system is applied no planned dialogue will take place as the ratings of managers/supervisors and peers are considered to be objective representations of work behaviour. It should be accentuated that on average two to three peer raters are used when an administrative rating systems is applied (Rasch 2004; Van Hooft et al. 2006). This study has shown that this number of peer raters leads to unsatisfying reliability levels and poor supervisor-peer correlations. Ratings as objective representations of work behaviour are much less emphasized in the personal growth system. The realization of an open dialogue should lead to a personal approach of competence development and a learning culture that should stimulate employees to reflect on personal strengths and weaknesses. In both systems supervisors/managers and peers assess the employees' level of competence, mostly using just a few items, constructed within an appraisal system (Arnold, Silvester, Cooper, Robertson and Burnes 2005; Blanchard and Thacker 2007). The input for both systems is comparable but only the personal growth system uses a planned dialogue to discuss differences and relate them to the rater's personal perspective. During the application of the administrative system the central focus of a dialogue is on goal setting. The judgment of supervisors/managers plays a major role in both systems, 360 degree feedback is very frequently applied in both systems to enhance reliability levels (George 1994; Mani 2002; Rasch 2004). The administrative system can be seen as a bureaucratic system, following strict administrative procedures, giving managers and SHRM/SHRD professionals concrete figures to analyze and control the current levels of competence as well as monitor progress. The use of single item multi-source ratings for feedback on behavioural change is very popular when the administrative system is used, with major emphasis on supervisor ratings (George 1994; Drenth 1998; Rasch 2004; Mani 2002). Few suggested procedures to systematically monitor progress seem to be available in either system (Mani 2002; Jellema 2003).

The deployment of three raters could be useful in the personal growth system, where high variability is accepted and differences in perspective should be used to explain different points of view to each other. Correlations of 0.35 (three raters) or 0.4 (four raters) should be considered too low for professional use. More than six peer raters are needed to reach a correlation above 0.45 between supervisor and peer ratings. The lack of dialogue concerning differences in ratings should be considered as a serious disadvantage of the administrative system. This is especially the case when the HRD policy focuses on personal qualities, because unreliable ratings on personal qualities evoke strong negative emotions (Brett and Atwater 2001). The same conclusion can be drawn for the use of six peer raters combined with a supervisor rating to effectuate a correlation higher than 0.45.

The use of 10 raters for a reliable 360 degree peer assessment and development system for competence development does not seem realistic in the daily practice of the work flow.

Organizational values create interesting opportunities describing major differences between the administrative system and the personal growth system. Key values of the administrative system are control and coordination by using data collected by strict administrative rules. Another key value of the administrative system is a strong result orientation. The underlying principles of the personal growth system are rooted in the development of potential capacities and the sharing of values of personnel. Lack of values in the organizational culture to discuss differences in perspective could be due to competing values of the organizational culture (Quinn 1991). Quinn (1991) argues that stimulating HRM/HRD activities will enhance the flexibility of organizations but will decrease the possibility of administrating important organizational processes. For organizations to be innovative the dominant influence of bureaucratic procedures has to be reduced, otherwise hidden talents and tacit knowledge cannot be developed or used. It is easy to imagine that the preferences of HRD professionals to use an administrative system for competence development are based on administrative, procedural cultural values. The same holds for the lack of values of the personal growth system to be result oriented. The values of the personal growth system such as potential development, cooperation and shared values compete with result oriented values in organizations (Quinn 1991). Differences in perspective based on differences in values easily lead to value conflicts (Quinn 1991). The dialogue to discuss differences in perspective demands quite excellent communicative leadership skills. Solving value conflicts is considered to be a very challenging leadership skill.

The model of competing values by Quinn (1991) seems to explain why receiving 360 degree feedback for administrative purposes has a lower popularity and could lead to more leniency than when used for personal growth reasons (London and Beatty 1993). The development of talents and potential of employees cannot be associated with the underlying values of an administrative system. This could undermine the intrinsic motivation of employees. This might be one of the reasons that some authors state that 360 degree feedback should be used for development purposes only (Dalessio 1998; Van Velsor 1998; Lepsinger and Lucia 2007). An important aspect of feedback on the development of competences, with personal qualities as developmental goals, is that personality traits and values form an essential part of an employee's identity. Unsolved differences concerning competing professional values increase the chance of escalating value conflicts (Quinn 1991). Andrew's (1997) research illustrated that shared values by managers and employees are important for ratings used for appraisal systems, especially if quality and the capacity to learn are important for the job. Studies have proven that value conflicts or greatly perceived heterogeneity in organizational values undermines group cohesion and causes stress, decreasing employee's performance and leading to output and quality problems (Dansereau and Alutto 1990; Bouckenhooghe, Buelens, Fontaine and Vanderheyden 2005). Guest (1998) shows that cooperation between management and subordinates will suffer seriously if the implicit expectations and communication between subordinates and management are damaged by value conflicts. Guest uses the terminology: 'breaking the psychological contract'. It is easy to understand that the feelings of being misjudged on values or personality traits cause serious problems concerning this psychological contract. This is especially the case for organizations delivering services, or organizations that depend strongly on a very high quality standard of products/services (Guest 1998). If one still wants to use an administrative system for competence management one should at least design a system in such a way that professional statistical analyses confirm the reliability and validity of dimensions and items. However, Drenth's (1998) work has shown that this is not common practice in the field of HRM/HRD.

Another possible solution would be to use validated instruments, but this would bring along extra financial costs and the necessity for the implementation of these instruments in the administrative system.

The personal growth system is based on values that state that the talents and potential of employees should be developed, leading to flexible knowledge management. A disadvantage of these values could be that monitoring and measuring HRD progress is easily neglected or gets very little emphasis. This effect could be reduced by systematically measuring feedback on HRD training effects by open and honest feedback. Another advantage of applying the personal growth system could be that it creates creative uneasiness. Creative uneasiness is an important condition for organizations to be innovative and to make use of talents and knowledge (Kessels 1996), however, it must be in balance with the necessary rest needed for the implementation of innovative changes. The essence of Kessels (1996) work is that real innovative power cannot be standardized in procedures, it has to be linked to important aspects of the organizational culture, like meta-cognitions, that are used to evaluate the latent learning mechanisms of the organization. Perhaps one can conclude that a personal growth system, that offers flexibility and stimulates employee's talents, should be preferred above an administrative system, when one supervisor and two or three peer raters are used. It could be misleading to rely on administrative data because it seems to be systematically collected and measurements are easily accessible. As the administrative system hardly creates possibilities for an open dialogue to discuss rating differences, chances are present that unreliable ratings are considered to be representative measures of real work behaviour, leading to moral and distrust problems. Another important reason, not to use only one supervisor rating and two or three peer ratings on competence development for administrative purposes, is related to the fact that feedback is important for the effectiveness of the development of competences (Hackman and Oldman 1975, 1980; Klein 1989; Houston 1990; Latham and Locke 1991; Brett and Atwater 2001). But the feedback must be reliable and trustworthy. The frequency of feedback should be decreased after employees have reached a satisfying level of competence development (Houston 1990). If these criteria are not fulfilled, feedback will not stimulate performance concerning competence development and will decrease intrinsic motivation. It seems to be quite evident that intrinsic motivation is necessary for competence development. Nobody will work on values and personality traits only because his manager wants him to.

Acknowledgement

The authors would like to express their gratitude to John Hayes for his additional editorial work.

References

Anderson, G., and Viswesvaran, C. (1998), 'An Update of the Validity of Personality Scales in Personal Selection: A Meta Analysis of Studies Published after 1992,' Paper presented at the 13th Annual Conference of the Society of Industrial and Organisational Psychology, Dallas.

Andrews, H. (1997), 'TQM and Faculty Evaluation: Ever the Twain Shall Meet?' Report No.BBB30994, Los Angeles, CA: ERIC Clearinghouse for Community Colleges (ERIC Document Reproduction Service No. ED 408 004).

Antonioni, D. (1996), 'Designing an Effective 360-Degree Appraisal Feedback Process,' *Organizational Dynamics*, 25, 2, 24–38.

Arnold, J., Silvester, J., Cooper, C.L., Robertson, I.T., and Burnes, B. (2005), *Work Psychology Understanding Human Behaviour in the Workplace*, Harlow, UK: Pearson Education Limited.

Arthur, W. Jr., Bennet, W. Jr., Edens, P.S., and Bell, S.T. (2003), 'Effectiveness of Training in Organizations: A Meta–Analysis of Design and Evaluation Features,' *Journal of Applied Psychology*, 88, 2, 234–245.

Arvey, R.D., and Murphy, K.R. (1998), 'Performance Evaluation in Work Settings,' *Annual Review of Psychology*, 49, 141–168.

Atkins, P.W.B., and Wood, R.E. (2002), 'Self Versus Others' Ratings as Predictors of Assessment Center Ratings: Validation Evidence for 360-Degree Feedback Programs,' *Personnel Psychology*, 55, 871–904.

Atwater, L., Roush, P., and Fischthal, A. (1995), 'The Influence of Upward Feedback on Self- and Follower Ratings of Leadership,' *Personnel Psychology*, 48, 1, 35–59.

Banks, C.G., and Murphy, K.M. (1985), 'Toward Narrowing the Research–Practice Gap in Performance Appraisal,' *Personnel Psychology*, 38, 335–345.

Barrick, M.R., and Mount, M.K. (1991), 'The Big Five Personality Dimensions and Job Performance: A Meta Analysis,' *Personnel Psychology*, 44, 1–26.

Barrick, M.R., Mount, M.K., and Judge, T.A. (2001), 'Personality and Performance at the Beginning of the New Millennium: What do we Know and Where do we go Next?' *Personality and Performance*, 9, 1/2, 9–29.

Bassi, L.J., and Van Buren, M.E. (1999), *The 1999 ASTD State of the Industry Report*, A supplement to Training and Development Magazine, 18, 414–432.

Beehr, T.A., Ivanitskaya, L., Hansen, C.P., Erofeev, D., and Gudanowski, D.M. (2001), 'Evaluation of 360–Degree Feedback Ratings: Relationships with Each Other and with Performance and Selection Predictors,' *Journal of Organizational Behaviour*, 22, 7, 775–788.

Berghenegouwen, G.J., and Mooijman, E.A.M. (2010), *Strategisch Opleiden en Leren in Organisaties* [Strategic Learning and Development in Organizations], Groningen: Kluwer.

Bernardin, H.J., and Pence, E.C. (1980), 'Effects of Rater Training: Creating New Response Sets and Decreasing Accuracy,' *Journal of Applied Psychology*, 65, 60–66.

Bettenhausen, K.L., and Fedor, D.B. (1997), 'Peer and Upward Appraisals: A Comparison of Their Benefits and Problems,' *Group and Organization Management*, 22, 236–263.

Blanchard, P.N., and Thacker, J.W. (2007), *Effective Training, Systems, Strategies, and Practices*, Englewood Cliffs, NJ: Pearson, Prentice Hall.

Bracken, D.W. (1994), 'Straight Talk About Multi–Rater Feedback,' *Training and Development*, 48, 9, 44–51.

Bracken, D.W., Timmreck, C.W., Fleenor, J.W., and Summers, L. (2001), '360 Feedback from Another Angle,' *Human Resource Management*, 40, 1, 3–20.

Brett, J.F., and Atwater, L.E. (2001), '360–Degree Feedback: Accuracy, Reactions and Perceptions of Usefulness,' *Journal of Applied Psychology*, 86, 5, 930–942.

Bretz, R.D. Jr., Milkovich, G.T., and Read, W. (1992), 'The Current State of Performance Appraisal Research and Practice: Concerns, Directions, and Implications,' *Journal of Management*, 18, 321–352.

Bouckenhooge, D., Buelens, M., Fontaine, J., and Vanderheyden, K. (2005), 'The Prediction of Stress by Values and Value Conflict,' *The Journal of Psychology*, 139, 4, 369–382.

Church, A.H., and Bracken, D.W. (1997), 'Advancing the State of the Art of 360–Degree Feedback: 'Guest Editors' Comments on the Research and Practice of Multi-Rater Assessment Methods,' *Group and Organization Management*, 22, 149–191.

Conway, J.M., and Huffcutt, A.I. (1997), 'Psychometric Properties of Multisource Performance Ratings: A Meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings,' *Human Performance*, 10, 331–360.

Dalessio, A. (1998), 'Using Multisource Feedback for Employee Development and Personnel Decisions,' in *Performance Appraisal: State of the Art in Practice*, ed. J.W. Smither, San Francisco, CA: Jossey-Bass, pp. 278–330.

Dansereau, F., and Alutto, J.A. (1990), 'Level of Analysis Issues in Climate and Culture Research,' in *Organizational Climate and Culture*, ed. B. Schnieder, San Francisco, CA: Jossey Bass, pp. 193–236.

De Gruijter, D.N.M., and Van der Kamp, L.J.Th. (2008), *Statistical Test Theory for the Behavioural Sciences*, Boca Raton, FL: Chapman and Hall.

Deming, W.E. (1986), *Out of the Crisis*, Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Deming, W.E. (1992), Presentation given in Phoenix, AZ. Pegasus Communications.

Drenth, P.J.D. (1998), 'Personnel Appraisal,' in *Handbook of Work and Organizational Psychology*, eds. P.J.D. Drenth, H. Thierry and C.J. De Wolff, Hove, UK: Psychology Press, pp. 59–88.

Fahr, J.L., Cannella, A.A., and Bedeian, A.G. (1991), 'Peer Ratings: The Impact of Purpose on Rating Quality and User Acceptance,' *Group and Organization Studies*, 16, 367–385.
Fried, Y., and Ferris, G.R. (1987), 'The Validity of the Job Characteristics Model: A Review and Meta Analysis,' *Personnel Psychology*, 40, 287–322.
Furnham, A. (2008), *Personality and Intelligence at Work. Exploring and Explaining Individual Differences at Work*, London: Routledge.
George, V. (1994), 'Performance Appraisal in an Academic Library: A Case Study,' Paper presented at the International Conference on TQM and Academic Libraries, Washington, DC.
Gerber, B. (1995), 'Does Training Make a Difference? Prove it!' *Training*, 18, 27–34.
Gray, G. (2002), 'Performance Appraisals Don't Work,' *Industrial Management*, 44, 15–17.
Greguras, G.J., and Robie, C. (1998), 'A New Look at Within–Source Interrater Reliability of 360–Degree Feedback Ratings,' *Journal of Applied Psychology*, 83, 960–968.
Guest, D. (1998), 'Is the Psychological Contract Worth Taking Seriously?' *Journal of Organizational Behaviour*, 19, 649–664.
Gwynne, P. (2002), 'How Consistent are Performance Review Criteria?' *MIT Sloan Management Review*, 43, 15.
Hackman, J.R., and Oldham, G.R. (1975), 'Development of the Job Diagnostic Survey,' *Journal of Applied Psychology*, 60, 159–170.
Hackman, J.R., and Oldham, G.R. (1980), *Work Redesign*, Reading, MA: Addison–Wesley.
Hamel, G., and Prahalad, C.K. (1994), *Competing for the Future*, Boston, MA: Harvard Business School Press.
Harris, M.M. (1994), 'Rater Motivation in the Performance Appraisal Context: A Theoretical Framework,' *Journal of Management*, 20, 737–756.
Hoffman, C.C., Nathan, B.R., and Holden, L.M. (1991), 'A Comparison of Validation Criteria: Objective Versus Subjective Performance Measures and Self-Versus Supervisor Ratings,' *Personnel Psychology*, 44, 3, 601–619.
Holland, J.L. (1997), *Making Vocational Choices: A Theory of Vocational Personalities and Work Environment* (3rd ed.), Odessa, FL: Psychological Assessment Resources Inc.
Houston, R. (ed.) (1990), *Handbook of Research on Teaching*, New York: MacMillan.
Hox, J.J. (2002), *Multilevel Analysis: Techniques and Applications*, Mahwah, NJ: Lawrence Erlbaum Associates.
Ilgen, D.R., and Feldman, J.M. (1983), 'Performance Appraisal: A Process Focus,' in *Research in Organizational Behaviour*, eds. L. Cummings and B. Staw, Greenwich, CT: JAI, pp. 141–197.
Jellema, F. (2000), 'Toepassing van 360-graden feedback in Nederlandse organisaties [Use of 360–degree feedback in Dutch organizations],' *Opleiding and Ontwikkeling*, 13, 21–25.
Jellema, F. (2003), 'Measuring Training Effects: The Potential of 360 Degree Feedback,' Doctoral Dissertation, University of Twente, Organisational Psychology & HRD.
Kenny, D.A., Albright, L., Malloy, T.E., and Kashy, D.A. (1994), 'Consensus in Interpersonal Perception: Acquaintance and the Big Five,' *Psychological Bulletin*, 116, 245–258.
Kessels, J.W.M. (1996), *Succesvol Ontwerpen (Successful Designing)*, Deventer: Kluwer Bedrijfswetenschappen.
Klein, H.J. (1989), 'An Integrated Control Theory Model of Work Motivation,' *Academy of Management Review*, 14, 150–172.
Landy, F.J., and Farr, J.L. (1983), *The Measurement of Work Performance: Methods, Theory, and Applications*, New York: Academic Press.
Lassiter, D. (1996), 'A User Guide to 360-Degree Feedback,' *Performance and Instruction*, 35, 5, 12–15.
Latham, G.P., and Locke, E.A. (1991), 'Self Regulation Through Goals Setting,' *Organisational Behaviour and Human Decision Processes*, 73, 753–772.
Lepsinger, R., and Lucia, A. (1997), *The Art and Science of 360-Degree Feedback*, San Francisco, CA: Pfeiffer.
London, M., and Beatty, R. (1993), '360–Degree Feedback as a Competitive Advantage,' *Human Resource Management*, 32, 2/3, 353–372.
London, M., and Smither, J.W. (1995), 'Can Multi–Source Feedback Change Perceptions of Goal Accomplishment, Self-Evaluations, and Performance–Related Outcomes? Theory–based Applications and Directions for Research,' *Personnel Psychology*, 48, 803–840.
Longenecker, C.O., Sims, H.P., and Gioia, D.A. (1987), 'Behind the Mask: The Politics of Employee Appraisal,' *Academy of Management Executive*, 1, 183–193.

Mani, B. (2002), 'Performance Appraisal Systems, Productivity, and Motivation: A Case Study,' *Public Personnel Management*, 31, 141–159.

Martineau, J. (1998), 'Using 360–Degree Surveys to Assess Change,' in *Maximizing the Value of 360–Degree Feedback*, ed. W. Tornow, San Fransisco, CA: Jossey–Bass, pp. 217–248.

McCarthy, A.M., and Garavan, T.N. (1999), 'Developing Self–Awareness in the Managerial Career Development Process: The Value of 360–Degree Feedback and the MBTI,' *Journal of European Industrial Training*, 23, 9, 437–445.

McCrae, R.R., and Costa, P.T. (1989), 'The Structure of Interpersonality Traits: Wiggin's Circumplex and the Five-Factor Model,' *Journal of Personality and Social Psychology*, 55, 586–595.

McCrae, R.R., and John, O.P. (1992), 'An Introduction of the Five Factor Model and its Applications,' *Journal of Personality*, 60, 175–215.

McEnery, J.M., and Blanchard, P.N. (1999), 'Validity of Multiple Ratings of Business Student Performance in a Management Simulation,' *Human Resource Development Quarterly*, 10, 2, 155–172.

Mount, M.K., Judge, T.A., Scullen, S.E., Sytsma, M.R., and Hezlett, S.A. (1998), 'Trait, Rater and Level Effects in 360-Degree Performance Ratings,' *Personnel Psychology*, 51, 557–576.

Murphy, K.R., and Cleveland, J.N. (1991), *Performance Appraisal: An Organizational Perspective*, Boston, MA: Allyn and Bacon.

Nunnally, J.C. (1978), *Psychometric Theory* (2nd ed.), New York: McGraw–Hill.

Pollack, D., and Pollack, L. (1996), 'Using 360-Degree Feedback in Performance Appraisal,' *Public Personnel Management*, 25, 4, 507–528.

Quinn, R.E. (1991), *Beyond Rational Management, Mastering the Paradoxes and Competing Demands of High Performance*, San Francisco, CA: Jossey-Bass.

Rasbash, J., Steele, F., Browne, W., and Prosser, B. (2004), *A User's Guide to MLwiN Version 2.0*, London: Institute of Education.

Rasch, G. (2004), 'Employee Performance Appraisal and the 95=5 Rule,' *Community College Journal of Research and Practice*, 28, 407–414.

Robbins, S.P. (2001), *Organizational Behaviour*, Englewood Cliffs, NJ: Prentice Hall.

Robinson, D., and Robinson, J. (1989), *Training for Impact: How to Link Training to Business Needs and Measure the Results*, San Fransisco, CA: Jossey Bass.

Roch, S.G., and McNall, L.A. (2007), 'An Investigation of Factors Influencing Accountability and Performance Ratings,' *The Journal of Psychology*, 141, 5, 499–523.

Salgado, J.F. (1997), 'The Five Factor Model of Personality and Job Performance in the European Community,' *Journal of Applied Psychology*, 82, 30–43.

Scullen, S.E., Mount, M.K., and Judge, T.A. (2003), 'Evidence of the Construct Validity of Developmental Ratings of Managerial Performance,' *Journal of Applied Psychology*, 88, 50–66.

Scholtes, P. (1999), 'Performance Appraisal: Book Review,' *Personnel Psychology*, 52, 177–181.

Society for Human Resource Management and Personnel Decisions International (2000), Performance Management Survey, http://www.shrm.org/searchcenter/Pager/Results.aspx?k=%20performance%20management%20survey

Toegel, G., and Conger, J.A. (2003), '360-degree Assessment: Time For Reinvention,' *Academy of Management Learning and Education*, 2, 297–311.

Van Hooft, E.A.J., van Flier, H. vander, and Minne, M.R. (2006), 'Construct Validity of Multi–Source Performance Ratings: An Examination of the Relationship of Self-, Supervisor-, and Peer-Ratings with Cognitive and Personality Measures,' *International Journal of Selection and Assessment*, 14, 25–81.

Van Velsor, E. (1998), 'Designing 360-Degree Feedback to Enhance Involvement, Self Determination, and Commitment,' in *Maximizing the Value of 360-Degree Feedback*, ed. W. Tornow, San Francisco, CA: Jossey-Bass, pp. 149–195.

Viswesvaran, C., Ones, D.S., and Schmidt, F.L. (1996), 'Comparative Analysis of the Reliability of Job Performance Ratings,' *Journal of Applied Psychology*, 81, 5, 557–574.

Waldman, D.A., and Atwater, L.E. (1998), *The Power of 360–Degree Feedback: How to Leverage Performance Evaluations for Top Productivity*, Houston, TX: Gulf Publishing Company.

Walker, A.G., and Smither, J.W. (1999), 'A Five–Year Study of Upward Feedback: What Managers Do With Their Results Matters,' *Personnel Psychology*, 52, 2, 393–423.

Walton, J. (1999), *Strategic Human Resource Development*, Essex: Pearson Education Limited.

Woehr, D.J., and Huffcut, A.I. (1994), 'Rater Training for Performance Appraisal: A Quantitative Review,' *Journal of Occupational and Organisational Psychology*, 65, 189–205.
Yammarino, F., and Atwater, L. (1997), 'Do Managers See Themselves As Others See Them?' *Organizational Dynamics*, 25, 4, 35–44.

Appendix

In this paper we study the reliability of the multi-source rating of one person, derived by aggregating (average or sum-score) a number of ratings obtained from members of a group this person is part of. Assuming independence for the ratings of the individual group members, but taking into account differences in shared context provided by membership of the same group, a reliability estimate can be obtained by computing

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n} + \hat{\sigma}_v^2}, \tag{1}$$

where σ_e^2 is the estimated variance between ratings of the same individual, σ_u^2 is the estimated variance across the aggregated ratings of the individuals within each group, σ_v^2 is the estimated variance of the average aggregated rating between groups, and n is the number of ratings for which the multi-source rating is derived.

Equation (1) expresses an approach from generalizability theory: the coefficient ρ is also known as the stepped-up intraclass correlation, but for this case the coefficient contains an additional variance component correcting for group membership (cf. De Gruijter and Van der Kamp 2008, p. 53–55). According to De Gruijter and Van der Kamp (2008) 'the coefficient is the generalizability counterpart of the reliability coefficient (α), its size giving information on the accuracy with which comparisons between persons can be made.

Estimates for ρ can easily be obtained from the variance component estimates of an unconditional three-level hierarchical model and the corresponding number of ratings per individual.